CollegeBoard
connect to college success™

Research Report
No. 2005-3

Evaluating SAT® II: Mathematics IC Items in the SAT I Population

Jinghua Liu, Fred Schuppan, and Michael E. Walker

www.collegeboard.com

College Board Research Report No. 2005-3

Evaluating SAT® II: Mathematics IC Items in the SAT I Population

Jinghua Liu, Fred Schuppan, and Michael E. Walker

College Entrance Examination Board, New York, 2005

Jinghua Liu is a measurement statistician at Educational Testing Service.

Fred Schuppan is an assessment specialist II at Educational Testing Service.

Michael E. Walker is a lead measurement statistician at Educational Testing Service.

The College Board: Connecting Students to College Success

The College Board is a not-for-profit membership association whose mission is to connect students to college success and opportunity. Founded in 1900, the association is composed of more than 4,700 schools, colleges, universities, and other educational organizations. Each year, the College Board serves over three and a half million students and their parents, 23,000 high schools, and 3,500 colleges through major programs and services in college admissions, guidance, assessment, financial aid, enrollment, and teaching and learning. Among its best-known programs are the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®). The College Board is committed to the principles of excellence and equity, and that commitment is embodied in all of its programs, services, activities, and concerns.

For further information, visit www.collegeboard.com.

Additional copies of this report (item #040481376) may be obtained from College Board Publications, Box 886, New York, NY 10101-0886, 800 323-7155. The price is $15. Please include $4 for postage and handling.

Printed in the United States of America.

Contents

Abstract

This study explored whether the addition of the items with more advanced math content to the SAT Reasoning Test™ (SAT®) would impact test-taker performance. Two sets of SAT math equating sections were modified to form four subforms each. Different numbers of items with advanced content, taken from the SAT II: Mathematics Level IC Test (Math IC), were embedded in the subforms. The number of Math IC items in a 25-item subform ranged from zero to six. The subforms were spiraled to obtain approximately equal numbers of examinees of roughly equal ability for all subforms. The Math IC items turned out to be more difficult than the SAT math items they replaced; therefore, there was a negative relationship between mean performance and number of Math IC items in the subforms. To examine possible indirect context effects of Math IC items, the 19 common SAT items across the subforms within a set were examined. When this was done, interset differences in performance disappeared. This finding supports the notion that test-taker performance is not affected by the mere presence of Math IC items. Rather, the effects of these items appear to be linked directly to the difficulty level of the items.

Introduction

The new SAT Reasoning Test (SAT) will be administered for the first time in March 2005. The new SAT math section will include topics typically taught in third-year high school math. In addition to covering content from Pre-Algebra, Algebra I, and Geometry, the new SAT math section will also contain content from Algebra II. To align the SAT further with classroom practice, quantitative comparisons will be eliminated.

It is important to gather preliminary data about how the new Algebra II content items will perform in the SAT population. The current SAT item pool does not contain any of these items. The SAT II: Mathematics Level IC Test (Math IC), however, does contain some of these items. The Math IC assesses students' understanding of the mathematics commonly taught in American high schools in three years of college preparatory mathematics (two years of Algebra and one year of Geometry). Therefore, Math IC items similar to those that will be included in the new SAT math section were administered to the SAT population to explore how this population would fare on these advanced-level items.

The study was carried out in the fall of 2002, within several of the fall SAT administrations. The study involved three phases. During Phase 1, a total of 12 Math IC items were embedded within different SAT math variable sections. The number of Math IC items included in any given 25-item section varied from zero to six. During Phase 2, 60 Math IC items were embedded in variable sections, six items per section. In Phase 3, the same 60 items were readministered in another SAT administration to collect data from a slightly different population from the previous population. The purpose of Phases 2 and 3 was mainly to collect pretest information for building prototypes of the new math section that would be administered in the spring 2003 new SAT field trial. However, Phase 1 was concerned with the effects of embedded Math IC items on test-taker performance. Thus, this paper focuses on the results of Phase 1.

Method

Participants

Data collection was carried out within a fall 2002 SAT administration. Two sets of subforms containing Math IC items were spiraled among the test-takers. Because the subforms were embedded in the variable sections, the Math IC items did not count toward test-takers' reported scores. Test-takers who took any of these subforms were considered participants of the study.

Materials

Two SAT math equating sections, with 25 items each, were modified to create two sets of four subforms each. The subforms in each set differed in the number of Math IC items they contained. For each set, the four subforms were created as follows:

- As a control, Subform 1 did not contain any Math IC items in the equating section (25 regular SAT items and no Math IC items). Thus, Subform 1 was identical to the original equating set;
- Subform 2 replaced two of the SAT items with Math IC items, so that 23 of the items were identical to those in Subform 1 (23 SAT items and 2 Math IC items);
- Subform 3 replaced two of the SAT items in Subform 2 with two additional Math IC items, so that 21 of the items in Subform 3 were identical to those in Subform 1 (21 SAT items and 4 Math IC items);
- Subform 4 replaced two of the SAT items in Subform 3 with two additional Math IC items, so that Subform 4 had 19 items in common with Subform 1 and 6 Math IC items (19 SAT items and 6 Math IC items).

To summarize, 12 Math IC items were embedded in the two equating sets with 6 items each. For each of the two sets, Subform 1 was the intact SAT math equating set. This subform was the parent form for Subforms 2 through 4. All four subforms had 19 SAT math items in common. Furthermore, each subform had 23 items in common

with its predecessor. The subforms did not overlap across the two sets. All of the subforms were spiraled during the administration, resulting in approximately equivalent groups taking each subform.

At the time the study was conducted, content specifications for the new SAT math section were not yet available. The chosen Math IC items represented a best guess about what content may appear on the new test. Of the 12 items, 9 contained content that was not tested in the current SAT math section. The remaining three items contained more formal mathematical language than that used on the current test, although the content of the items is currently covered on the SAT math section. Although items similar to the 12 chosen for this study may appear on the new SAT math section, the items should not be considered representative of all content changes.

Results

Samples

Overall, 46,088 test-takers took the four subforms in Set 1, and 46,060 test-takers took the four subforms in Set 2. Among those taking Set 1, 19,828 were male and 26,260 were female. For the Set 2 sample, 19,691 were male and 26,369 were female. Information on race/ethnicity was collected as well. In each set, there were around 3,000 test-takers in each of the Asian American, African American, and Hispanic groups, and there were about 23,000 white test-takers.

Comparison of Item Statistics for Math IC and SAT Math Items

Table 1 gives item statistics, including item difficulty (equated delta) and item discriminating power (r-biserial), for the six Math IC items in each set and for the SAT math items they replaced. Each Math IC item was put in the same position as the SAT math item it replaced. Thus we can make a direct comparison between each Math IC item and the SAT math item in the same position in the subform.

As can be seen from Table 1, almost all Math IC items were more difficult than the corresponding SAT items that they replaced, except item #5 in Set 1. The mean delta of the six Math IC items in Set 1 was 15.4, much higher than the mean delta of 13.1 for the six SAT math items. The same was true of the Math IC items in Set 2, with an average delta of 14.9 as compared with 12.8 for the SAT math items.

The r-biserial comparisons showed that Math IC items had lower discriminating power than the SAT items on average. In Set 1, the Math IC items had a mean r-biserial of 0.50, lower than the mean r-biserial of the corresponding SAT math items (0.56). The Math IC items in Set 2 also manifested a lower correlation with the total subform, with an average r-biserial of 0.56 compared to a mean r-biserial of 0.59 for SAT items.

Table 1

Comparison of Item Statistics for Math IC Items and the SAT Math Items They Replaced

		SAT I: Math items		*SAT II: Math IC items*	
Set 1	***Sequence #***	***Equated Delta***	***r-biserial***	***Equated Delta***	***r-biserial***
	5	9.2	0.56	9.2	0.54
	22	15.7	0.55	20.0	0.49
	7	11.5	0.53	14.3	0.50
	23	15.9	0.57	19.4	0.26
	13	11.4	0.61	14.2	0.55
	16	14.7	0.51	15.3	0.63
	Average	13.1	0.56	15.4	0.50
Set 2	***Sequence #***	***Equated Delta***	***r-biserial***	***Equated Delta***	***r-biserial***
	5	8.1	0.31	12.4	0.63
	21	16.7	0.62	18.7	0.46
	23	17.0	0.68	20.5	0.56
	4	7.5	0.62	10.5	0.54
	18	15.1	0.74	16.7	0.59
	7	12.4	0.58	10.8	0.59
	Average	12.8	0.59	14.9	0.56

The lower average correlation with the subform for Math IC items reflects a stronger negative relationship between r-biserials and deltas among these items than among the SAT math items in the sets. That is, more difficult Math IC items tend to exhibit lower correlations with the criterion (i.e., the total subform). A contributing factor is that the correlation of the individual item score with the total test score becomes lower as the item becomes harder (or easier). The use of biserial as opposed to Pearson correlation coefficients corrects somewhat for this tendency, however. In addition, the Math IC items come from a test that measures a somewhat different construct than does the SAT math section. Thus, it is not surprising that the Math IC items exhibit slightly lower correlations with the total score than do the SAT math items. In any event, the average difference in r-biserials between the two sets was only slight in this sample of items.

Comparison of Section Statistics Across Subforms

Table 2 provides information on the statistics for each of the two sets by subform. The subforms are listed by the number of Math IC items they contain: from no items to six items. The mean equated delta and r-biserial are listed for each form.

Data in Table 2 indicate that the difficulty level of the section increased when more Math IC items were

Table 2

Comparison of Statistics Across Subforms

Number of Math IC Items	*Equated Delta*		*r-biserial*	
	Mean	*SD*	*Mean*	*SD*
Set 1				
0	12.34	3.46	0.54	0.10
2	12.50	3.64	0.55	0.09
4	12.70	3.86	0.53	0.10
6	12.97	3.91	0.53	0.11
Set 2				
0	12.47	3.37	0.54	0.13
2	12.86	3.46	0.54	0.12
4	12.99	3.46	0.54	0.11
6	13.06	3.56	0.54	0.10

embedded in the subform. This finding was consistent with the item statistics discussed above and shown in Table 1: The individual Math IC items were more difficult than the SAT math items they replaced. There was actually an attempt to control item difficulty between original and replacement items when the tests were assembled. However, the difficulty statistics for the Math IC items were based on a more able population (i.e., the SAT Subject Test population) than the SAT population on which the SAT math item difficulty statistics were based. Thus, a Math IC item with a delta statistic of 12 would be more difficult in general than an SAT item with a delta statistic of 12. As a result, the section difficulty naturally increased as the more difficult Math IC items were added.

The data in Table 2 may suggest at first glance that the SAT math section will necessarily become more difficult as more Math IC items are added. This is not the case. The SAT math section will continue to be built to very rigid specifications, in terms of both item difficulty and item discrimination. Thus we can expect average performance for the total group on the new SAT math section to be the same as the average performance on the current SAT-M. Table 2 does make clear, however, that we cannot gain insight into the effects of adding Math IC items simply by examining average performance on the subforms. There are other ways to elucidate the effects of adding Math IC items. These include the examination of differential item functioning, item impact, subform impact, and item context effects. Each of these is treated in turn below.

Comparison of Item Fairness Statistics

One means of assessing the appropriateness of Math IC items for the new SAT math section is to examine item fairness statistics, in particular differential item functioning (DIF) and item impact statistics. An item will exhibit impact if a larger percentage of people in one intact group (ethnic/race or gender group) answer an item correctly than in another intact group. An item will exhibit DIF if, for individuals with the same total score on the SAT math section, those in a particular ethnic/race or gender group have a greater chance of answering the item correctly than those in another ethnic/race or gender group.[1]

Although these statistics are collectively referred to as fairness statistics, the existence of a large DIF or impact statistic does not necessarily mean that an item is unfair. An item will exhibit DIF if it measures a construct other than the major construct measured by the test, and if groups differ on this second construct to a greater or lesser degree than they differ on the major construct. An item will exhibit impact if ethnic/race or gender groups differ on the construct of interest. Examination of DIF and impact statistics for Math IC items will help to determine whether these items measure a construct different from what the SAT math section measures; and whether greater group differences can be expected on these items than on the SAT math section.

Table 3 presents the Mantel-Haenszel Delta-DIF (MH D-DIF)[2] and impact statistics for all of the Math IC items, compared to the SAT math items being replaced. The criterion was total SAT math score. Table 3 shows that in Set 1, Math IC items produced lower average MH D-DIF than the SAT math items for the male/female (0 vs. 0.36), and white/Asian American (0.13 vs. 0.36) comparisons on average, and produced slightly higher MH D-DIF for the white/African American (0.61 vs. 0.18), and white/Hispanic (0.13 vs. -0.01) comparisons, favoring the African American and Hispanic groups. No items were flagged for C-DIF.[3]

Data in Set 2 indicated a slightly different pattern from those in Set 1. Math IC items produced smaller MH D-DIF than SAT math items for the male/female group comparison. For all of the ethnic/race group comparisons, Math IC items produced larger MH D-DIF than the SAT math items. Note that the Math IC items produced positive MH D-DIF values on average, favoring the focal groups (female, African American, Hispanic,

1. In practice, minority groups (Asian American, African American, Hispanic, and American Indian examinees) are compared to white examinees; and females are compared to males.
2. These statistics are based on the Mantel-Haenszel method of DIF detection. The statistics are scaled so that a D-DIF statistic of 1 corresponds roughly to a one-delta difference in the difficulty of an item for the two populations.
3. For the SAT Program, items for which the absolute values of MH D-DIF are greater than 1.5 and are statistically significantly greater than 1.0 are placed in Category C, the most extreme category of DIF. These items are screened from the item pool and are not used in tests.

Table 3

Comparison of Item Fairness Statistics

		Male/Female		*White/African Am.*		*White/Hispanic*		*White/Asian Am.*	
	Seq.	*DIF*	*Impact*	*DIF*	*Impact*	*DIF*	*Impact*	*DIF*	*Impact*
Set 1									
SAT Math	5	0.94	0.01	0.28	-0.14	-0.27	-0.11	1.14	0.06
	22	0.38	-0.05	0.88	-0.10	0.68	-0.07	1.25	0.20
	7	0.13	-0.05	0.36	-0.23	-0.12	-0.13	0.01	0.05
	23	-0.24	-0.10	-0.74	-0.19	-0.28	-0.13	-0.26	0.08
	13	0.64	-0.02	0.17	-0.21	-0.47	-0.18	0.16	0.07
	16	0.31	-0.05	0.14	-0.18	0.4	-0.10	-0.15	0.18
Average		0.36	-0.04	0.18	-0.18	-0.01	-0.12	0.36	0.11
Math IC	5	0.30	-0.02	0.77	-0.10	0.32	-0.08	0.58	0.05
	22	-0.04	-0.05	0.47	-0.05	0.00	-0.04	0.22	0.08
	7	-0.13	-0.09	0.24	-0.17	-0.20	-0.14	-0.65	0.04
	23	-0.38	-0.05	0.34	-0.03	0.06	-0.02	-0.35	0.04
	13	-0.11	-0.10	0.86	-0.13	0.14	-0.11	0.37	0.12
	16	0.36	-0.07	1.01	-0.14	0.45	-0.09	0.60	0.14
Average		0.00	-0.06	0.61	-0.10	0.13	-0.08	0.13	0.08
Set 2									
SAT Math	5	-0.18	-0.03	-0.06	-0.11	-0.19	-0.07	-0.46	-0.01
	21	0.99	-0.03	0.31	-0.14	-0.13	-0.12	0.47	0.15
	23	-0.28	-0.10	0.19	-0.14	0.03	-0.10	0.30	0.13
	4	-0.57	-0.04	0.01	-0.14	0.13	-0.07	0.47	0.02
	18	-1.30	-0.19	-1.32	-0.32	-0.23	-0.19	-0.49	0.09
	7	0.31	-0.05	0.78	-0.16	0.05	-0.14	0.31	0.09
Average		-0.17	-0.07	-0.02	-0.17	-0.06	-0.12	0.10	0.08
Math IC	5	-0.14	-0.08	0.83	-0.20	0.82	-0.10	0.47	0.08
	21	0.29	0.03	0.08	-0.08	0.32	-0.04	0.53	0.11
	23	0.22	0.09	0.23	-0.04	0.58	-0.02	1.48	0.14
	4	0.05	-0.05	-0.07	-0.18	-0.31	-0.13	-0.35	0.01
	18	-0.42	-0.10	0.20	-0.14	0.26	-0.08	0.24	0.12
	7	0.20	-0.04	0.49	-0.19	0.09	-0.12	-0.08	0.03
Average		0.03	-0.03	0.29	-0.14	0.29	-0.08	0.38	0.08

and Asian American). The Math IC item in slot #23 was detected as displaying positive C-DIF for the white/Asian American comparison in one subform. No other Math IC items were detected as C-DIF items.

Table 3 also shows impact statistics for each item and for the average. For all comparison groups in both Set 1 and Set 2, the average impact of the Math IC items was either equal to or lower than the average impact produced by SAT math items. Neither impact nor DIF results indicate an increased overall impact with the inclusion of these six Math IC items.

Comparison of Subform Fairness Statistics

In Table 3, DIF and impact statistics were listed separately for each of the Math IC items and the SAT math items they replaced. Recall that most of these items appeared in more than one subform. We can also examine the average DIF and impact statistics by subform. Often, DIF analyses are performed on the items in a test form using the total test score as the stratification criterion. In these cases, the average DIF value is near zero as an artifact of the procedure. When the criterion includes the analyzed

items, then, the average DIF value is uninformative. In our situation, these averages are informative because the stratification criterion for DIF and impact is the total SAT math score, not the subform score.

Average MH D-DIF and impact statistics for each subform are shown in Table 4. These averages are taken across all 25 items in a subform (remember that 19 of these items are the same across all subforms within a set). For each gender and ethnic/race comparison, for both Sets 1 and 2, the average impact remained fairly constant across subforms with different numbers of Math IC items. For both sets, a general trend may be seen in the data such that the subforms with greater numbers of Math IC items tend to favor the ethnic/race focal group (i.e., Asian Americans, African Americans, Hispanics) more and the reference group (i.e., whites) less. The addition of Math IC items in Set 1 tends to favor males over females; whereas the addition of Math IC items in Set 2 tends to favor females over males. In general, though, there are no great differences among the average DIF statistics across subforms.

The results in Table 4 are strictly a function of the items used in the study and may be discerned from the data in Table 3. Table 4 is useful, however, because it illustrates the overall effect of including items that may exhibit a certain amount of DIF. The table also shows no clear-cut pattern of DIF or impact associated with the inclusion of more Math IC items.

Table 4

Comparison of Section Fairness Statistics

Set 1	*Male/Female*							
Number of Math IC Items	*MH D-DIF*				*Impact*			
	Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
0	0.19	0.53	-1.13	1.28	-0.04	0.03	-0.13	0.01
2	0.10	0.41	-0.80	0.82	-0.05	0.03	-0.15	-0.01
4	0.18	0.44	-0.69	0.97	-0.04	0.03	-0.13	0.01
6	0.11	0.50	-1.40	0.97	-0.05	0.03	-0.13	0.00
	White/African American							
Number of Math IC Items	*MH D-DIF*				*Impact*			
	Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
0	0.07	0.45	-0.74	0.89	-0.16	0.06	-0.25	-0.04
2	0.06	0.55	-1.18	1.13	-0.15	0.06	-0.25	-0.05
4	0.07	0.38	-0.82	0.78	-0.15	0.07	-0.24	-0.03
6	0.19	0.60	-0.92	1.01	-0.14	0.07	-0.26	-0.02
	White/Hispanic							
Number of Math IC Items	*MH D-DIF*				*Impact*			
	Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
0	-0.08	0.47	-1.16	0.77	-0.11	0.04	-0.18	-0.04
2	-0.08	0.46	-1.12	0.88	-0.12	0.04	-0.18	-0.04
4	-0.03	0.52	-1.41	1.13	-0.09	0.05	-0.16	0.00
6	0.02	0.55	-1.20	0.97	-0.08	0.04	-0.17	-0.01
	White/Asian American							
Number of Math IC Items	*MH D-DIF*				*Impact*			
	Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
0	0.14	1.00	-2.35	1.36	0.08	0.07	-0.08	0.20
2	-0.15	0.85	-2.34	1.47	0.06	0.06	-0.04	0.20
4	-0.04	0.89	-1.75	1.71	0.07	0.07	-0.07	0.21
6	-0.14	0.85	-1.78	1.32	0.06	0.07	-0.06	0.17

Table 4 (Continued)

Comparison of Section Fairness Statistics

Set 2	*Male/Female*							
Number of Math IC Items	*MH D-DIF*				*Impact*			
	Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
0	-0.07	0.63	-1.30	1.12	-0.07	0.05	-0.19	0.01
2	-0.07	0.67	-1.31	1.15	-0.07	0.05	-0.18	0.02
4	-0.01	0.61	-1.26	1.22	-0.07	0.05	-0.19	0.02
6	0.01	0.54	-0.86	0.95	-0.06	0.05	-0.15	0.01
	White/African American							
Number of Math IC Items	*MH D-DIF*				*Impact*			
	Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
0	0.14	0.47	-1.32	1.20	-0.17	0.07	-0.32	-0.05
2	0.14	0.47	-0.81	0.96	-0.16	0.07	-0.29	-0.03
4	0.14	0.47	-0.48	0.98	-0.16	0.07	-0.28	-0.05
6	0.19	0.45	-0.55	1.35	-0.16	0.07	-0.29	-0.03
	White/Hispanic							
Number of Math IC Items	*MH D-DIF*				*Impact*			
	Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
0	-0.02	0.45	-0.95	0.97	-0.11	0.05	-0.02	-0.21
2	0.12	0.36	-0.47	0.86	-0.09	0.04	-0.16	-0.01
4	0.15	0.54	-1.22	1.18	-0.11	0.05	-0.25	-0.02
6	-0.01	0.56	-1.30	1.25	-0.10	0.05	-0.23	-0.01
	White/Asian American							
Number of Math IC Items	*MH D-DIF*				*Impact*			
	Mean	*SD*	*Min*	*Max*	*Mean*	*SD*	*Min*	*Max*
0	0.14	0.53	-1.02	1.01	0.08	0.06	-0.01	0.19
2	0.02	0.57	-1.06	0.93	0.05	0.05	-0.03	0.14
4	0.18	0.56	-0.78	1.34	0.07	0.06	-0.01	0.19
6	0.16	0.72	-1.30	1.80	0.07	0.07	-0.02	0.20

Comparison of Subform Performance

Table 5 presents descriptive data on how students did on each set of subforms with different numbers of Math IC items. The table presents data for all 25 items in a subform. As illustrated in Tables 1 and 2, the subforms differed in difficulty because no information was available to control for the difficulty of the Math IC items and the SAT math items they replaced. Thus, we see a general pattern that average scores are lower on subforms with more Math IC items. This phenomenon should not be taken as a function of the number of Math IC items but rather as a function of the difficulty of the subforms. In addition to showing the average performance by set and subform, the table breaks down performance by ethnicity/race and gender.

Total Group

The first section of Table 5 gives the total group performance on the 25-item subforms. The data show, not surprisingly, that the average score decreased as the number of Math IC items increased. This trend corresponded to the average item difficulty for each subform reported in Table 2. The lower average scores were associated with the more difficult subforms.

Gender Groups

The performance of gender groups is shown in the second and third sections of Table 5. Both male and female groups displayed a pattern of means similar to that of the total group. Test-takers' performance was negatively affected by the more difficult Math IC items. A finding of more interest to us here concerns the relationship between the averages for males and females on each subform. One could conjecture that the inclusion of

Table 5

Comparison of Test-Taker Performance on the 25-Item Subforms by Gender and Ethnicity/Race

No. Math IC Items	*Set 1*			*Set 2*		
	N	*Mean*	*SD*	*N*	*Mean*	*SD*
Total Group						
0	12,129	13.78	5.53	12,101	13.68	5.94
2	11,794	13.55	5.47	11,826	13.08	5.88
4	11,281	13.04	5.31	11,287	12.79	5.89
6	10,884	12.66	5.34	10,846	12.74	5.72
Male						
0	5,203	14.52	5.55	5,150	14.79	5.91
2	5,140	14.31	5.55	5,014	14.20	5.93
4	4,907	13.75	5.35	4,886	13.97	5.88
6	4,578	13.40	5.44	4,641	13.75	5.76
Female						
0	6,926	13.23	5.45	6,951	12.87	5.83
2	6,654	12.96	5.33	6,812	12.25	5.69
4	6,374	12.49	5.21	6,401	11.89	5.74
6	6,306	12.13	5.20	6,205	11.98	5.57
Asian American						
0	988	16.82	5.76	1,010	16.62	6.16
2	934	16.04	5.91	985	15.77	6.13
4	867	15.56	5.69	877	15.66	5.93
6	881	15.06	5.96	881	15.70	6.33
African American						
0	876	9.72	5.23	851	9.25	5.25
2	887	9.57	5.12	838	8.57	5.42
4	782	9.28	5.02	810	8.46	5.37
6	748	9.13	4.92	796	8.41	5.20
Hispanic						
0	783	10.80	5.28	771	10.66	5.74
2	781	10.44	5.21	811	10.29	5.78
4	773	10.31	5.37	759	9.57	5.83
6	709	10.24	5.24	721	9.77	5.52
White						
0	6,082	14.29	5.06	6,113	14.13	5.46
2	5,890	14.06	4.97	5,908	13.52	5.46
4	5,585	13.41	4.81	5,638	13.21	5.47
6	5,536	13.01	4.96	5,380	13.14	5.23

more advanced mathematics content could exacerbate gender differences on the subforms. The data do not support such an assertion, however. For Set 1 and for Set 2, the standardized difference between male and female performance[4] remained fairly constant, even as the number of Math IC items increased. For Set 1, the standardized difference was about -0.24 for all subforms. For Set 2, the average standardized difference across subforms was about -0.33. In each case, as evidenced in the table, males outperformed females.

Ethnic/Race Groups

The last four sections of Table 5 present the information on ethnic/race groups. Again, for each of the ethnic/race groups, the mean score on the subform decreased as the number of Math IC items increased. For Set 2, the standardized differences remained fairly constant across subforms for the Asian American/white (average difference = 0.42), African American/white (average difference = -0.82), and Hispanic/white (average difference = -0.58) comparisons. For Set 1, however, the standardized differences decreased in absolute magnitude as the number of Math IC items increased: For the Asian American/white comparison, the standardized difference ranged from 0.46 on the subform with no Math IC items to 0.38 for the form with 6 Math IC items; for the African American/white comparison, the standardized difference went from -0.82 to -0.73; and for the Hispanic/white comparison, the difference decreased from -0.63 to -0.52.

The relationship between number of Math IC items and group membership can be more easily seen by means of graphs. Plots of mean score differences for Set 1 under different Math IC conditions for total group, gender, and ethnic/race groups can be found in Figures 1 through 3. Figure 1 shows the consistent pattern of decrease in test scores as more Math IC items are added. Figure 2, showing mean performance on subforms by gender, also illustrates that the decrease in performance across subforms is similar for both males and females.

Figure 3 illustrates how the results differed by ethnic/race group. The mean decreases in performance with added Math IC items were more pronounced for the Asian American group, which exhibited the highest mean score on all subforms. The change in means across subforms was smallest for African American examinees, whose overall mean was the lowest of all ethnic/race groups. This difference across ethnic/race groups coincides with the decrease in standardized differences discussed previously.

Because the difficulty of Math IC items was not controlled, the number of Math IC items in a subform was confounded with the difficulty of that form. Therefore, differences in the 25-item performance across subforms cannot be attributed unambiguously to the inclusion of Math IC items. Context effects of including different numbers of Math IC items in subforms, however, can be investigated. To do this, the

4. Here, the standardized difference is defined as the mean for males minus the mean for females, divided by the standard deviation for the total group.

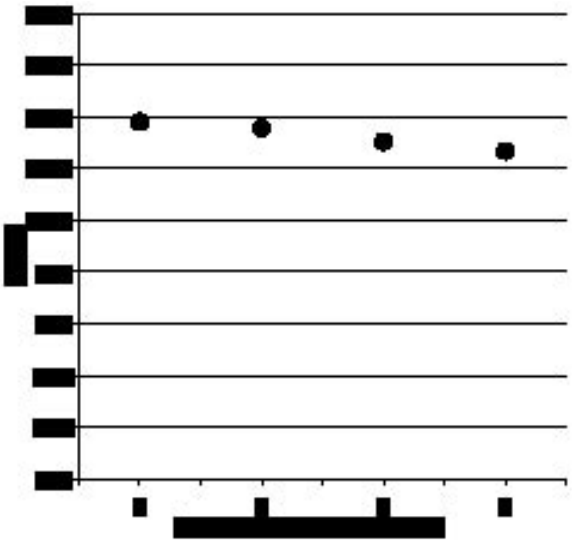

Figure 1. Means for total group, 25-item subform.

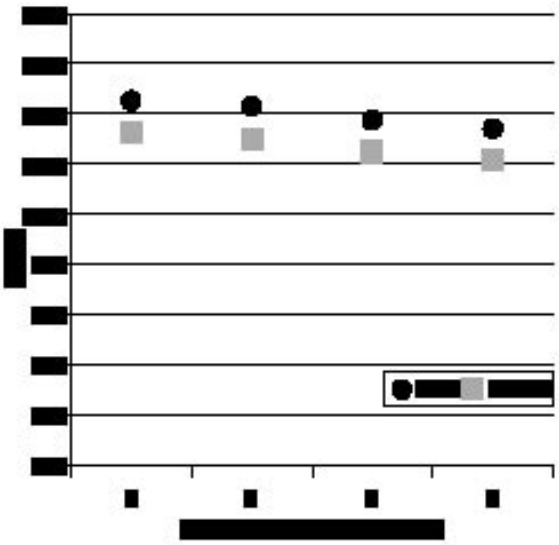

Figure 2. Means by gender, 25-item subform.

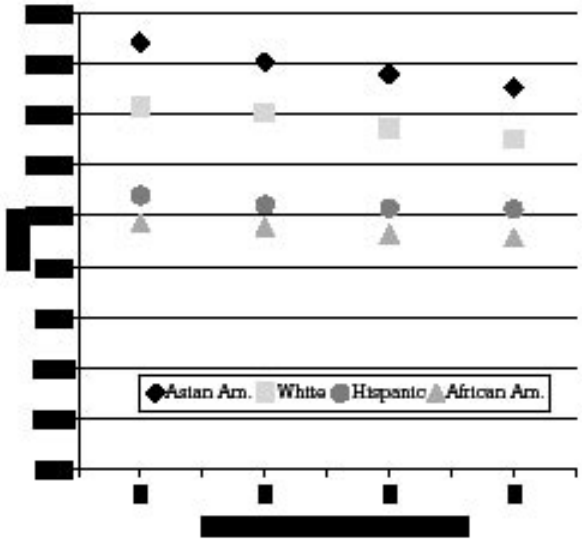

Figure 3. Means by ethnicity, 25-item subform.

19 common items across subforms were examined to determine if there was any difference in performance on these items depending upon the number of Math IC items included with them. These analyses are reported in the following pages.

Comparison of Performance on the 19 Common Items Across Subforms

As discussed above, the difficulty of different subforms varied due to the greater difficulty of the Math IC items used in this study as compared with the SAT math items they replaced. In practice, the new SAT math section will continue to be built to rigorous item difficulty specifications. Thus, there should be no similar decrease in test performance because of more difficult Math IC items being added.

One possible consequence of adding Math IC items to the SAT math section could be that the mere presence of these items could contribute to decreased performance on the other items in the test. For example, Math IC items could take longer to complete than SAT math items, thus leaving less time to answer other items. Such context effects would not manifest themselves necessarily in the item difficulty indices for the Math IC items. Instead, increased numbers of omitted items might be seen, or smaller percentages of examinees may reach the items at the end of the test.

To check for possible context effects associated with adding Math IC items to the subforms, the 19 items common to all subforms were examined. Because these items appeared in the same sequence in all subforms, the items were the same in all respects except for the number of Math IC items in the subforms. Mean differences among subforms that were not due solely to performance on the Math IC items themselves (i.e., were not reflected in the item difficulty) would still be evident after the Math IC items were removed.

Plots of mean performance on the 19 common items in Set 1 for the total, gender, and ethnic/race groups can be found in Figures 4 through 6. Figure 4 shows that the means for the four subforms were very close to each other. The means did not decrease with increased numbers of Math IC items, as they did in Figure 1. Similarly, the trends evident in Figures 2 and 3 are not apparent in Figures 5 and 6.

Table 6 displays the means on the 19 common items for the subforms in Sets 1 and 2. The data in Table 6 also show that the means were very close to each other across the subforms. As shown in Figures 5 and 6, gender and ethnic/race differences persisted even after the removal of the Math IC items, but the differences across subforms with different numbers of Math IC items were no longer evident. Formal tests of significance were conducted to

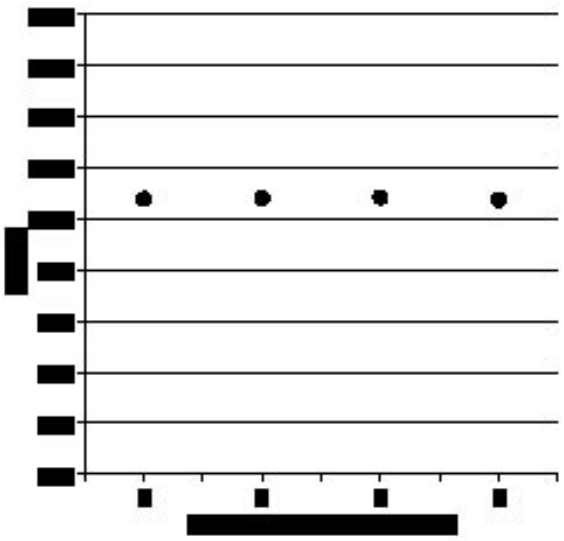

Figure 4. Means for total group, 19 common items.

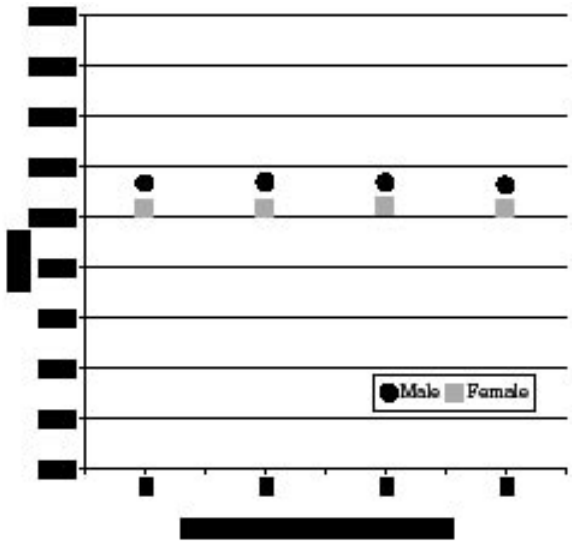

Figure 5. Means by gender, 19 common items.

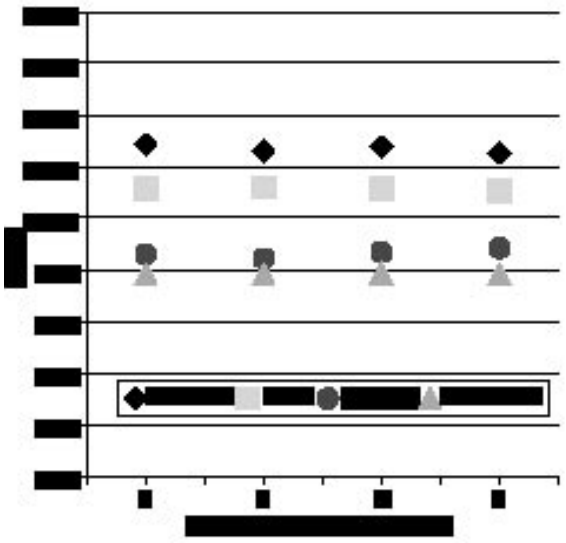

Figure 6. Means by ethnicity, 19 common items.

Table 6

Comparison of Test-Taker Performance on the 19 Common Items by Gender and Ethnicity/Race

No. Math IC Items	*Set 1*			*Set 2*		
	N	*Mean*	*SD*	*N*	*Mean*	*SD*
Total Group						
0	12,129	10.76	4.14	12,101	10.62	4.59
2	11,794	10.80	4.21	11,826	10.45	4.49
4	11,281	10.82	4.18	11,287	10.48	4.57
6	10,884	10.73	4.19	10,846	10.41	4.53
Male						
0	5,203	11.33	4.15	5,150	11.44	4.57
2	5,140	11.37	4.27	5,014	11.26	4.51
4	4,907	11.34	4.16	4,886	11.36	4.52
6	4,578	11.24	4.22	4,641	11.21	4.53
Female						
0	6,926	10.33	4.08	6,951	10.02	4.51
2	6,654	10.36	4.11	6,812	9.85	4.38
4	6,374	10.42	4.14	6,401	9.81	4.49
6	6,306	10.35	4.12	6,205	9.81	4.43
Asian American						
0	988	12.85	4.36	1,010	12.80	4.71
2	934	12.59	4.56	985	12.26	4.70
4	867	12.74	4.49	877	12.56	4.53
6	881	12.53	4.63	881	12.54	4.89
African American						
0	876	7.82	3.95	851	7.27	4.24
2	887	7.83	4.02	838	7.03	4.28
4	782	7.87	4.11	810	7.14	4.31
6	748	7.81	4.06	796	6.98	4.23
Hispanic						
0	783	8.61	3.99	771	8.32	4.53
2	781	8.44	4.11	811	8.36	4.51
4	773	8.69	4.31	759	7.97	4.54
6	709	8.83	4.27	721	8.03	4.45
White						
0	6,082	11.14	3.80	6,113	10.98	4.22
2	5,890	11.19	3.82	5,908	10.83	4.15
4	5,585	11.12	3.75	5,638	10.82	4.24
6	5,536	11.03	3.84	5,380	10.77	4.16

investigate the existence of any differences among the conditions.

Overall Math IC Effects

Tables 7 and 8 show the results of the analyses of variance comparing performance differences on the subforms for Sets 1 and 2, respectively. For each set of subforms, two

ANOVAs were conducted. The first analysis examined gender differences and the second analysis looked at ethnic/race group differences. Both analyses are shown in each table.

The first thing to notice when comparing the two analyses for each set is that the sample size for the ethnic/race group analyses is about 30 percent smaller than the sample size for the gender group analyses. The reason for this is that several examinees did not indicate an ethnicity or race. These individuals were excluded from the second analysis, but not from the first. Because the first analysis included the larger sample, it was used to assess the overall effect of embedded Math IC items.

As can be seen from Table 7, the differences in mean scores among the four subforms for Set 1 were not significant (p = .34). Thus, the data offered no evidence that the inclusion of Math IC items affected performance on other SAT math items. In other words, any difference in performance on a subform was due to the difficulty of the Math IC items, not the mere presence of those items.

Table 8, on the other hand, indicates a statistically significant effect of number of Math IC items included in the subform. Follow-up analysis using the Least Significant Difference procedure showed that the mean performance on the common items in the subform with no Math IC items was statistically significantly better than performance on the same 19 items in subforms with Math IC items. Table 6 reveals, however, that the greatest mean difference (between the 0 IC and the 6 IC subforms) is 0.21 out of 19 possible points. This value corresponds to a standardized difference of less than 0.05. In addition, the effect size coefficient (η^2) was very trivial, around 0.0003. Thus, although the effect was statistically significant, it was far from noteworthy or important.

Table 7

Analysis of Variance for Test-Taker Performance on the 19 Common Items Under Different Conditions: Set 1

Source of Variation	*SS*	*df*	*MS*	*F*	*p*	η^2
Gender Groups						
Between	10,570.41	7	1,510.06	87.72	0.000	
Items	57.32	3	19.11	1.11	0.344	
Gender	10,497.80	1	10,497.80	609.79	0.000	0.0131
Item × Gender	15.30	3	5.10	0.30	0.828	
Within	793,281.21	46,080	17.22	—		
Total	803,851.63	46,087	—	—		
Ethnic/Race Groups						
Between	58,787.46	15	3,919.16	251.56	0.000	
Items	25.80	3	8.60	0.55	0.647	
Ethnicity	58,586.81	3	19,528.94	1,253.49	0.000	0.102
Item × Ethnic	174.85	9	19.43	1.25	0.261	
Within	515,467.68	33,086	15.58	—		
Total	574,255.14	33,101	—	—		

Table 8

Analysis of Variance for Test-Taker Performance on the 19-Item Equating Section Under Different IC Conditions: Set 2

Source of Variation	*SS*	*df*	*MS*	*F*	*p*	η^2
Gender Groups						
Between	24,021.41	7	3,431.63	170.38	0.000	0.025245
Items	302.90	3	100.97	5.01	0.002	0.000326
Gender	23,671.37	1	23,671.37	1,175.31	0.000	0.024886
Item × Gender	47.15	3	15.72	0.78	0.505	
Within	927,515.60	46,052	20.14	—		
Total	951,537.01	46,059	—	—		
Ethnic/Race Groups						
Between	73,191.26	15	4,879.42	264.77	0.000	0.107035
Items	370.99	3	123.66	6.71	0.000	0.000607
Ethnicity	72,758.48	3	24,252.83	1,316.00	0.000	0.106470
Item × Ethnic	61.79	9	6.87	0.37	0.949	
Within	610,612.46	33,133	18.43	—		
Total	683,803.72	33,148	—	—		

Gender Effects

Tables 7 and 8 also give the ANOVA results comparing mean score performance between the gender groups. There was a statistically significant main effect of gender. For all subforms in both Sets 1 and 2, males outperformed females. This result was expected, given the typical difference between males and females on the current SAT math section. The average difference on the subforms was about 1 point (standardized difference of 0.23) for Set 1 and 1.4 points (standardized difference of 0.31) for Set 2. The gender differences were fairly consistent across subforms within a set, as evidenced by the nonsignificant gender by subform interaction. Thus, the presence of Math IC items in the subforms did not have a differential impact across genders on performance on the SAT math items.

Ethnic/Race Effects

The bottom sections of Tables 7 and 8 present the ANOVA results for the ethnic/race group comparisons. The main effect of the ethnic/race group was statistically significant for both item sets. Table 6 (and Figure 6) presents the directions of the differences among ethnic/race groups, with Asian American examinees scoring highest on average and African American examinees scoring lowest.

All differences were statistically significant. The largest separation in performance was between Asian American and African American groups, with an average difference of about 5.0 points out of 19 possible points (standardized difference = 1.2). The smallest difference was between African American and Hispanic groups, with an average difference of about 0.8 points (standardized difference = 0.2). As in the case of the gender differences, these results coincide with findings on the current SAT math section. Also as before, the interaction with subform was not statistically significant.

Discussion

The results of Phase 1 indicated that when more difficult Math IC items were embedded in SAT math subforms, test-taker performance declined. The decline in performance was directly related to the difficulty of the items that were embedded. A logical next phase for research would be to administer Math IC items and SAT math items of comparable difficulty and to look for differences in performance. Such a design was not possible for this study, however, because the study represented the first opportunity to administer Math IC items to the general SAT population and to obtain difficulty statistics on the SAT difficulty scale.

The prototypes administered in the 2003 field trial are built to SAT specifications, and they contain items with similar content to Math IC items. The field trial offers an opportunity to judge whether or not the content of Math IC items presents problems for test-takers. A logical supposition would be that there should be no difference in the function of Math IC items and SAT math items with the same difficulty statistics. If something about the Math IC item made it more difficult than its matched SAT math item, for example, then the difficulty rating of the Math IC item would reflect this difference. In that case, however, the two items would not be matched in difficulty. Thus it is hard to imagine many properties of Math IC items that would not be revealed in the item statistics and thus controlled by the test specifications.

One possible property that falls outside the item statistics involves length of time needed to answer the item. Even if item timing does not directly affect the item statistics, it will affect performance on other items and overall performance on the test. One might think of other such properties of Math IC items as well. This study examined these effects by studying examinee performance on the 19 common items across several subforms with varying numbers of Math IC items. If the context of embedded Math IC items had affected performance, then one would have expected to see differences in performance across conditions.

To the contrary, performance on the common items across conditions was virtually identical. Furthermore, there were no differential context effects across gender or ethnic/race subgroups. These findings give us confidence that there should be no adverse effects of embedding Math IC or similar items in the SAT math section, so long as the item difficulties are appropriate for the population and so long as the test continues to be built to the same specifications.


CollegeBoard
connect to college success™